Probabilistic Riemannian submanifold learning with wrapped Gaussian process latent variable models

Anton Mallasto
Department of Computer Science
University of Copenhagen

Søren Hauberg
DTU Compute
Technical University of Denmark

Aasa Feragen
Department of Computer Science
University of Copenhagen

Abstract

Latent variable models (LVMs) learn probabilistic models of data manifolds lying in an *ambient* Euclidean space. In a number of applications, a priori known spatial constraints can shrink the ambient space into a considerably smaller manifold. Additionally, in these applications the Euclidean geometry might induce a suboptimal similarity measure, which could be improved by choosing a different metric. Euclidean models ignore such information and assign probability mass to data points that can never appear as data, and vastly different likelihoods to points that are similar under the desired metric. We propose the wrapped Gaussian process latent variable model (WGPLVM), that extends Gaussian process latent variable models to take values strictly on a given ambient Riemannian manifold, making the model blind to impossible data points. This allows non-linear, probabilistic inference of low-dimensional Riemannian submanifolds from data. Our evaluation on diverse datasets show that we improve performance on several tasks, including encoding, visualization and uncertainty quantification.

1 INTRODUCTION

Unsupervised learning aims at modelling structure in unlabeled data, such as its geometry. Sometimes, information on this geometry is available through spatial constraints or a non-Euclidean metric, e.g. the data lives on a Riemannian manifold. Incorporating the known Riemannian manifold in a probabilistic model should improve model fit, and save us from learning what we already know. In this work, we study a probabilistic latent variable model that takes the geometry into account.

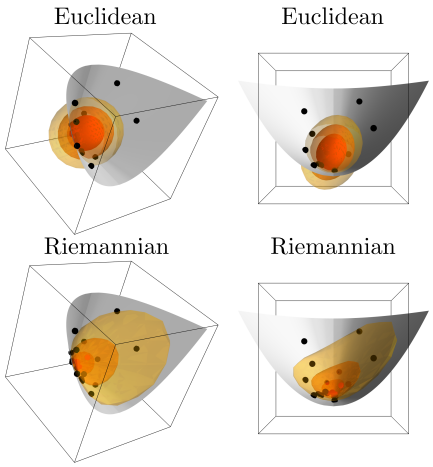

Figure 1: The ambient manifold $SPD(2)$ is the open subset *on the inside* of the visualized grey cone in the ambient Euclidean space $\mathbb{R}^3$. **Top row:** A Euclidean Gaussian distribution fitted to a set of $SPD(2)$ matrices (black dots) escapes *outside* of $SPD(2)$. **Bottom row:** The Riemannian Log-Euclidean metric yields a wrapped Gaussian distribution that remains inside $SPD(2)$, providing a better fit to the data. The colored trust regions are confidence regions of the (W)GDs.

Where do manifolds come from? Data points on a sphere are forced to have norm one, covariance matrices are symmetric and positive definite, and shapes do not depend on scale, rotation or placement. Enforcing such constraints or invariances, one replaces the ambient Euclidean space by an ambient *manifold*. The *ambient space* refers to the set of all those points, which the model views as possible data points. The constraints alter the shortest paths between data objects, giving rise to a *Riemannian metric*. Riemannian metrics can also be imposed by modelling choices; closeness under the Euclidean metric does not always express desired

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

similarity of data objects. These metrics can be learned from data (Hauberg et al., 2012) or imposed based on domain knowledge (Arsigny et al., 2006).

Euclidean probabilistic models on manifold data assign probability mass to impossible data points under spatial constraints. Furthermore, points that are similar under the chosen non-Euclidean metric can be assigned very different likelihoods, which can cause a poor fit to the data. Both issues affect especially the uncertainty estimates. These issues can be avoided by exploiting the Riemannian geometry in the model. Fig. 1 shows points in $SPD(2)$, the space of 2×2 symmetric positive-definite matrices, with fitted Euclidean and Riemannian models. The points outside the cone are not $SPD(2)$ matrices. Under the Log-Euclidean metric, which generalizes the log transform to matrices, elements on the boundary (in gray) lie infinitely far from interior points. The metrics, and hence the induced models, are vastly different. This results in the Riemannian model with an improved model fit.

Contributions. Motivated by these observations, we introduce the *wrapped Gaussian process latent variable model* (WGPLVM). This extends the Gaussian process latent variable model (GPLVM) to data on Riemannian manifolds by employing *wrapped Gaussian processes* (WGPs). Like the GPLVM, the WGPLVM defines a probabilistic model between elements in a lower dimensional *latent space* and the data, providing uncertainty estimates. As WGPs take values strictly on a given Riemannian manifold, the WGPLVM enforces known constraints and invariances, and accounts for modelling choices concerning the metric.

We demonstrate the WGPLVM on several different manifolds and tasks. We show that our method provides more efficient encoding of the original data compared to the Euclidean GPLVM, provides superior uncertainty estimates and better captures trends in the data, resulting in improved visualization results.

Related Literature. First, we discuss methods in *manifold learning*, which view data points as elements of a Euclidean space. Then, we discuss related work in *submanifold learning*, that works strictly on Riemannian manifolds. Note that some manifold learning methods can impose known geometry on the latent space. Models relying on kernels (e.g. the GPLVM and WGPLVM) can encode such structure on the latent space (Lin et al., 2017). This is different from imposing geometric constraints on the data space.

Manifold learning infers a low-dimensional manifold that captures the trend of given data. Classical algorithms (Belkin and Niyogi, 2003; Roweis and Saul, 2000; Tenenbaum et al., 2000) learn a low distortion projection from a data submanifold of the original,

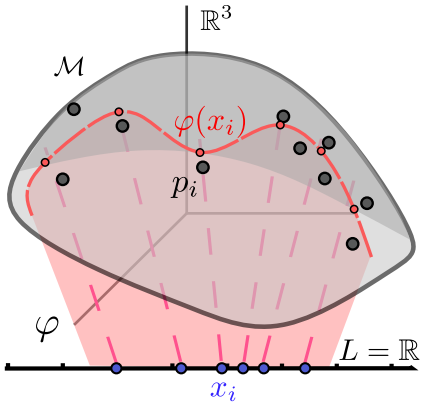

Figure 2: Illustration of submanifold learning.

Euclidean ambient space, onto a low-dimensional Euclidean space. Latent variable models (LVMs) (Goodfellow et al., 2014; Kingma and Welling, 2014; Lawrence, 2005) learn the reverse *latent embedding* from the latent space into the ambient space, associating each point in the latent space with an ambient space point. In the well-known *Gaussian process latent variable model* (GPLVM) (Lawrence, 2005), the latent embedding is a Gaussian process (GP) over the latent space, and hence learns not only a manifold embedding into $\mathbb{R}^n$, but also a model of its uncertainty. GPLVMs have inspired other LVMs (Lawrence and Moore, 2007; Titsias and Lawrence, 2010; Urtasun and Darrell, 2007), that all rely on Euclidean geometry. Urtasun et al. (2008) consider topologically constrained LVMs and Varol et al. (2012) consider GPLVMs with spatial constraints, where the constraints are enforced through slack variables and local linearization. Our method works intrinsically on the specific Riemannian manifold, taking the topology, spatial constraints and the Riemannian metric into account. Thus the WGPLVM falls into the category of submanifold learning.

Submanifold learning algorithms, illustrated in Fig. 2, aim to infer a model φ from a latent space L to a submanifold M (dashed red) of a known *ambient manifold* $\mathcal{M}$ of points that satisfy the constraints. The map φ associates the data $p_i \in \mathcal{M}$ (dark grey) with latent variables $x_i \in L$ (blue). *Principal geodesic analysis* (PGA) (Fletcher et al., 2004; Huckemann et al., 2010) estimates geodesic submanifolds, *Riemannian principal curves* (Hauberg, 2016) and *barycentric subspaces* (Pennec, 2015) estimate less constrained submanifolds. *Probabilistic PGA* (Zhang and Fletcher, 2013) introduces uncertainty by estimating probabilistic geodesic subspaces. The WGPLVM contributes non-geodesic, probabilistic learning of the submanifold from a prior model, allowing considerable flexibility compared to previous models.

Examples of manifold valued data include directional

statistics, which consider spherical data (Mardia and Jupp, 2009; Urtasun et al., 2006), covariance matrices as data objects in economics and computer vision (Tuzel et al., 2006; Wilson and Ghahramani, 2011) and in diffusion MRI or materials science (Batchelor et al., 2005; Fletcher and Joshi, 2004), and statistics of shape, which is of fundamental interest in computer vision (Freifeld and Black, 2012; Kendall, 1984). In each example, the common approach is to incorporate the Riemannian structure in the statistical analysis.

2 PRELIMINARIES

This section introduces the necessary preliminaries and notation. We first review Gaussian processes (GPs) and the Gaussian process latent variable model (GPLVM) (Lawrence, 2004). Next, we summarize the necessary concepts from Riemannian geometry. Subsequently, we review the wrapped Gaussian processes (WGPs) introduced by Mallasto and Feragen (2018), which form the cornerstone of the present work.

Gaussian processes. Let $\mathcal{N}(\mu, \Sigma)$ denote a multivariate Gaussian distribution (GD) with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$, and write the associated probability density function as $\mathcal{N}(v|\mu, \Sigma)$ for $v \in \mathbb{R}^d$. A *Gaussian process* (GP) is a collection f of random variables, so that any finite subcollection $(f(\omega_i))_{i=1}^N$ is jointly Gaussian, where $\omega_i \in \Omega$ are elements of the *index set*. Any GP f is uniquely characterized by

$$\begin{aligned} m(\omega) &= \mathbb{E}\left[f(\omega)\right], \\ k(\omega, \omega') &= \mathbb{E}\left[(f(\omega) - m(\omega))(f(\omega') - m(\omega'))^T\right], \end{aligned} \tag{1}$$

called the *mean function* m and *covariance function* k, denoted $f \sim \mathcal{GP}(m, k)$. For more about GPs and their applications, see Rasmussen (2004).

Gaussian process latent variable model. The Gaussian process latent variable model (GPLVM) is a GP-based dimensionality reduction technique, which aims to learn a probabilistic model relating elements in the low dimensional *latent space* $L \subseteq \mathbb{R}^{n'}$ to observed data $Y = \{y_i\}_{i=1}^N \subset \mathbb{R}^n$, with $n' < n$. The model approximates the manifold that Y lives on. The probabilistic model is computed by choosing a prior GP $f \sim \mathcal{GP}(m, k_\theta)$ with hyper-parameters $\theta \in \Theta$. The hyper-parameters are optimized with the *latent variables* $X = \{x_i\}_{i=1}^N \in L$ to maximize the log-likelihood

$$\begin{aligned} \log(\mathbb{P}(Y|X,\theta)) = &-\frac{nN}{2}\ln(2\pi) - \frac{n}{2}\ln|K_{X,\theta}| \\ &-\frac{1}{2}\mathrm{Tr}\left(K_{X,\theta}^{-1}YY^T\right), \end{aligned} \tag{2}$$

where $(K_{X,\theta})_{ij} = k_\theta(x_i, x_j)$, and X, Y denote the corresponding data matrices. Finally, we condition the optimal prior f on the chosen latent variables X and data Y, to yield the predictive distribution of the model. Note that any prediction $f(x)$ has support in the whole $\mathbb{R}^n$, thus ignoring any constraints or invariances.

In differential geometric terms, a GPLVM can be viewed to learn a stochastic *chart* for the approximate manifold on which the dataset Y lives.

Riemannian geometry. A *Riemannian manifold* is a *smooth manifold* M with a *Riemannian metric*, i.e. a smoothly varying inner product $g_p(\cdot,\cdot)$ on the tangent space T_pM at each $p \in M$, which induces a distance function d_M on M. Each (p, v) in the *tangent bundle* $TM = \bigcup_{p\in M}(\{p\} \times T_pM)$ defines a *geodesic* γ (locally shortest path) on M, so that $\gamma(0) = p$ and $\dot\gamma(0) = v$.

The Riemannian *exponential map* $\mathrm{Exp}\colon TM \to M$ is given by $(p, v) \mapsto \mathrm{Exp}_p(v) = \gamma(1)$, where γ is the geodesic corresponding to (p, v). The exponential Exp_p at p is a diffeomorphism between a neighborhood $0 \in U_p \subset T_pM$ and a neighborhood $p \in V_p \subset M$, which is chosen in a maximal way to preserve injectivity. The *logarithmic map* $\mathrm{Log}_p\colon V_p \to T_pM$ is characterized by the identity $\mathrm{Exp}_p(\mathrm{Log}_p(p')) = p'$. Outside of V_p, we use $\mathrm{Log}_p(p')$ to denote $v \in \mathrm{Exp}_p^{-1}(p')$ with a minimal norm, chosen in a *measurable* way. The complement of V_p in M is called the *cut-locus* at p, where unique geodesics cannot be defined. Multiple useful manifolds have empty cut-locus, so that $V_p = M$, including manifolds with non-positive curvature as well as the space of positive-definite symmetric matrices used below.

Let $\mathrm{Exp}_p(v) = q$ and $\gamma(t) = \mathrm{Exp}_p(tv)$. The differential $D_p\mathrm{Log}_p(q)$ (in some coordinate chart) is given by (see supplementary material for (Pennec, 2016))

$$D_p\mathrm{Log}_p(q) = \left(J_0(1)\right)^{-1} J_1(1), \tag{3}$$

where J_i are Jacobi fields solving the linear ordinary differential equation

$$\ddot J_i(t) + R(t)J_i(t) = 0, \tag{4}$$

with initial conditions $J_0(0) = 0$, $\dot J_0(0) = I_n$, and $J_1(0) = I_n$, $\dot J_1(0) = 0$. Here $R(t)$ is given by $R_{ij} = \langle \mathrm{Riem}_{\gamma(t)}(\dot\gamma(t), e_i(t))\dot\gamma(t), e_j(t)\rangle_{\gamma(t)}$ and $(e_1(t), ..., e_n(t))$ is an orthonormal basis for $T_{\gamma(t)}M$, defined by $e_1(0) = \frac{v}{\|v\|_2}$ and each $e_j(t)$ evolves through parallel transportation. Furthermore, $\mathrm{Riem}_{\gamma t}$ denotes the curvature tensor and I_n is the n-by-n identity matrix, where n is the dimension of the manifold. For a thorough exposition in Riemannian geometry, see (Do Carmo, 1992).

Let M_i be Riemannian manifolds with metrics g_i, exponential maps Exp^i and logarithmic maps Log^i for $i = 1, 2$. Then $M = M_1 \times M_2$ turns into a Riemannian manifold when endowed with the metric $g = g_1 + g_2$, which has the component-wise computed exponential map $\mathrm{Exp}_{(p_1,p_2)}((v_1, v_2)) = \left(\mathrm{Exp}^1_{p_1}(v_1), \mathrm{Exp}^2_{p_2}(v_2)\right)$. The logarithmic map Log on the product manifold is

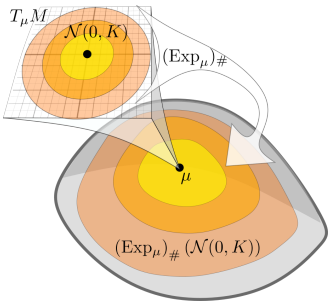

Figure 3: WGDs defined as a Gaussian $\mathcal{N}(0,K)$ in the tangent space $T_\mu M$ over the basepoint μ, which is pushed forward by the exponential map Exp_μ to M.

defined likewise.

Wrapped Gaussian distributions. Let (M,g) be an n-dimensional geodesically complete Riemannian manifold. Let ν be a measure on X and $f\colon X \to Y$ be a measurable map. We define the *push-forward* as $f_\# \nu(A) := \nu(f^{-1}(A))$ for any measurable set A in Y. A random point X on M follows a *wrapped Gaussian distribution* (WGD), if for some $\mu \in M$ and a symmetric positive definite matrix $K \in \mathbb{R}^{n\times n}$

$$X \sim \left(\mathrm{Exp}_\mu\right)_\# \left(\mathcal{N}(0,K)\right), \tag{5}$$

denoted by $X \sim \mathcal{N}_M(\mu,K)$. The WGD is thus defined by a GD $\mathcal{N}(0,K)$ in the tangent space $T_\mu M$, that is pushed-forward onto M by the exponential map Exp_μ (see Fig. 3). We call $\mu =: \mu_{\mathcal{N}_M}(X)$ the *basepoint* of X, and $K =: \mathrm{Cov}_{\mathcal{N}_M}(X)$ the *tangent space covariance*.

Two random points $X_i \sim \mathcal{N}_{M_i}(\mu_i, K_i)$, $i = 1,2$ are *jointly WGD*, if (X_1, X_2) is a WGD on the product manifold $M_1 \times M_2$, given by

$$(X_1,X_2) \sim \mathcal{N}_{M_1\times M_2}\left(\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}, \begin{pmatrix}K_1 & K_{12}\\ K_{21} & K_2\end{pmatrix}\right), \tag{6}$$

for some matrix $K_{12} = K_{21}^T$. Then, X_1 can be conditioned on X_2, resulting in a push-forward of a Gaussian mixture in $T_{\mu_1}M_1$ by the exponential map

$$X_1|(X_2 = p_2) \sim \left(\mathrm{Exp}_{\mu_1}\right)_\# \left(\sum_{v\in A} \lambda_v \mathcal{N}\left(\mu_v, K_v\right)\right), \tag{7}$$

where $A = \{v \in T_{\mu_2}M \mid \mathrm{Exp}_{\mu_2}(v) = p_2\}$ is the preimage of p_2. The means and covariance matrices of the Gaussian mixture components are given by

$$\mu_v = K_{12}K_2^{-1}v, \qquad K_v = K_1 - K_{12}K_2^{-1}K_{12}^T, \tag{8}$$

and the component weights are

$$\lambda_v = \frac{\mathcal{N}(v|\mathbf{0},K_2)}{\mathbb{P}\{A\}}, \quad \mathbb{P}\{A\} = \sum_{v\in A} \mathcal{N}(v|\mathbf{0},K_2). \tag{9}$$

Wrapped Gaussian processes. Wrapped Gaussian processes generalize GPs to Riemannian manifolds (Mallasto and Feragen, 2018). A collection f of random points on a Riemannian manifold M indexed over a set Ω is a *wrapped Gaussian process* (WGP), if every finite subcollection $(f(\omega_i))_{i=1}^N$ is jointly WGD on M^N. The functions

$$\begin{aligned} m(\omega) &= \mu_{\mathcal{N}_M}(f(\omega)),\\ k(\omega,\omega') &= \mathrm{Cov}_{\mathcal{N}_M}(f(\omega), f(\omega')), \end{aligned} \tag{10}$$

are called the *basepoint function* and the *tangent space covariance function* of f (also called as kernel of f), respectively. To denote such a WGP, we use the notation $f \sim \mathcal{GP}_M(m,k)$.

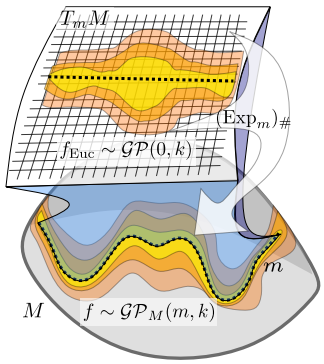

Figure 4: A WGP f can be viewed as defining a GP f_{Euc} in the tangent spaces $T_m M \subset M$ over the basepoint function, so that each marginal $f(x_i)$ is pushed-forward onto M by $(\mathrm{Exp}_{m(x_i)})_\#(f(x_i))$.

Formally, a WGP f can be viewed as a GP f_{Euc} on $T_m M \subset TM$, the family of tangent spaces over the basepoint function m. Then, the resulting GP is pushed forward to M using the Riemannian exponential map Exp_m over m to obtain the WGP, see Fig. 4.

3 WRAPPED GAUSSIAN PROCESS LATENT VARIABLE MODEL

We now introduce the *wrapped Gaussian process latent variable model* (WGPLVM) for data $P = \{p_i\}_{i=1}^N$ lying on an n-dimensional ambient Riemannian manifold $\mathcal{M}$. The goal of WGPLVM is to learn a lower-dimensional submanifold $M_{Pred} \subset \mathcal{M}$, where the data is assumed to reside. The WGPLVM model is a straight-forward generalization of the GPLVM model, where instead of GPs, we maximize the likelihood of our data combined with the latent variables under the WGPs that are suitable for the manifold context. The WGPLVM pipeline is illustrated in Fig. 5.

We consider a family of WGPs $f \sim \mathcal{GP}_{\mathcal{M}}(m, k_\theta)$ from the latent space L onto the ambient manifold $\mathcal{M}$, where $\theta \in \Theta$ are hyperparameters, that will be optimized over. The basepoint function m can be utilized to delocalize

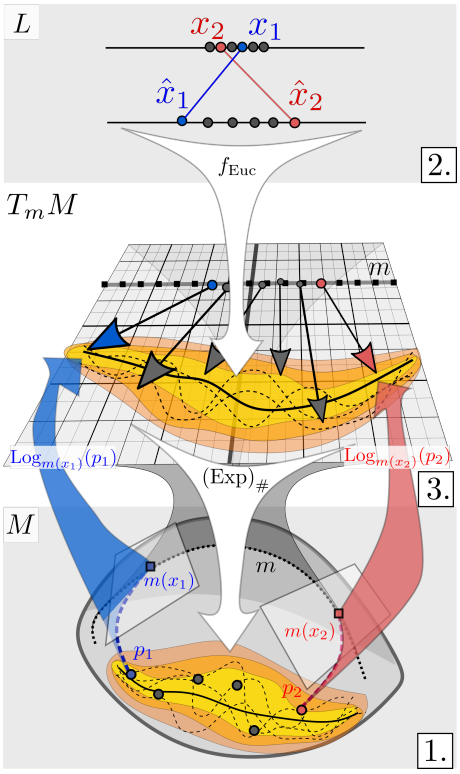

Figure 5: **The WGPLVM pipeline. 1.** The data $p_i \in \mathcal{M}$ (blue and red dots) is transformed to the tangent bundle by $p_i \mapsto \text{Log}_{m(x_i)}(p_i) \in T_{m(x_i)}\mathcal{M} \subset T_m\mathcal{M}$ along the prior basepoint function m (dotted black line) at initial latent variables x_i. **2.** A GPLVM is learned, yielding the latent variables $\hat{x}_i \in L$ and the GP f_{Euc} from L to the tangent bundle. **3.** The GP f_{Euc} is then pushed forward onto $\mathcal{M}$ by $(\text{Exp})_\#(f_{\text{Euc}})$, resulting in the predicted data submanifold.

the learning process in order to avoid distortions of the metric caused by linearization of the curved $\mathcal{M}$. The kernel k_θ affects how observations in different tangent spaces affect each other. For coherence, the kernel should be adapted to a smooth *frame* (a smoothly changing basis over m). Such a frame can e.g. be constructed by *parallel transporting* a basis along m.

The likelihood assigned by the prior f to a data point p with associated latent variable x is

$$\begin{aligned}\mathbb{P}\{p|x,\theta\} &= \sum_{v \in \text{Exp}_{m(x)}^{-1}(p)} \mathcal{N}(v|\mathbf{0}, K_{x,\theta}) \\ &\approx \mathcal{N}\left(\text{Log}_{m(x)}(p)|\mathbf{0}, K_{x,\theta}\right),\end{aligned} \tag{11}$$

where $(K_{x,\theta})_{ij} = k_\theta(x^i, x^j)$ and $x = (x^1, x^2, ..., x^n)$.

The approximation in Eq. (11) only takes into account the preimage of p with a minimal norm (and thus maximal likelihood), denoted by $\text{Log}_{m(x)}(p)$. The expression gives a lower bound for $\mathbb{P}\{p|x,\theta\}$, thus, maximizing the likelihood of $\text{Log}_{m(x)}(p)$ maximizes the lower bound for $\mathbb{P}\{p|x,\theta\}$. It also enforces the WGPLVM to prefer *local* models over ones that wrap considerably around the manifold. Note that, for manifolds with empty cut-locus (such as ones with non-positive curvature), the approximation in (11) is exact.

The objective function to be maximized is then the approximated log-likelihood

$$\begin{aligned}\ln\left(\mathbb{P}\{p|x,\theta\}\right) \approx &-\frac{dN}{2}\ln(2\pi) - \frac{d}{2}\ln|K_{x,\theta}| \\ &-\frac{1}{2}\text{Log}_{m(x)}(p)^T K_{x,\theta}^{-1}\text{Log}_{m(x)}(p),\end{aligned} \tag{12}$$

for which the gradient with respect to x is given by

$$\begin{aligned}&\frac{\partial}{\partial x_j}\ln\left(\mathbb{P}\{p|x,\theta\}\right) \approx \\ &-\frac{d}{2}\text{Tr}\left(K_{x,\theta}^{-1}\frac{\partial K_{x,\theta}}{\partial x_j}\right) \\ &-\frac{1}{2}\text{Log}_{m(x)}(p)^T K_{x,\theta}^{-1} D_{m(x)}\text{Log}_{m(x)}(p)\frac{\partial m}{\partial x_j}(x) \\ &-\frac{1}{2}\text{Log}_{m(x)}(p)^T\frac{\partial K_{x,\theta}^{-1}}{\partial x_j}\text{Log}_{m(x)}(p),\end{aligned} \tag{13}$$

The differential $D_{m(x)}\text{Log}_{m(x)}(p)$ can be computed using Jacobi fields as explained in expression (3), if no analytical expression exists.

Assuming that the data is i.i.d, the approximate log-likelihood for the data set P can be written using Eq. (12), by considering P as a single element of the product manifold P^N. This quantitity is then maximized by optimizing over the latent variables and the hyperparameters θ, resulting in the optimal latent variables $\hat{X}$ and hyperparameters $\hat{\theta}$ for the kernel.

The approximate submanifold can then be predicted at arbitrary latent variables X_{Pred}, by conditioning $\hat{f} \sim \mathcal{GP}_{\mathcal{M}}(m, k_{\hat{\theta}})$ on the data P with the associated latent variables $\hat{X}$ (using Eq. (7)). The conditional distribution will then be a non-centered GP $f_{\text{Euc}} \sim \mathcal{GP}(m_{\text{Euc}}, k_{\text{Euc}})$ defined on $T_m\mathcal{M}$ pushed forward by the exponential map (see Fig. 5), resulting in the predictive distribution $\varphi_{\text{pred}} \sim (\text{Exp}_{m(x)})_\#(f_{\text{Euc}})$. Then, the *mean prediction* is given by $\bar{\varphi}_{\text{pred}}(x) = (\text{Exp}_{m(x)})_\#(m_{\text{Euc}})(x))$.

In Eq. (7), if the preimage $\text{Exp}_{\mu_2}^{-1}(p_2)$ is not uniquely defined, the conditional distribution is approximated by using a preimage with minimal norm, as previously. This approximation is justified as the weights λ_v of the components of the Gaussian mixture decrease exponentially as $\|v\|_{p_2}$ increases.

The initial latent variables $X = \{x_i\}_{i=1}^N$ can be chosen strategically to aid optimization. We use *principal geodesic analysis* (PGA) (Fletcher et al., 2004) and *principal curves* (Hauberg, 2016). PGA is appropriate

when the data expresses a geodesic trend (analogy of linearity on Riemannian manifolds), which is not the case for the femur dataset, see Fig. 6 in Section 4.

The Computational complexity for the method is $\mathcal{O}(NL + N^3)$, where L is the cost of computing the Riemannian logarithm. This varies from manifold to manifold, but for example, in Section 4, the most expensive is $\mathcal{O}(d^3)$ for the Log-Euclidean metric on $d \times d$ symmetric, positive-definite matrices.

We provide a pseudo-algorithm for the method in the supplementary material.

4 EXPERIMENTS

The WGPLVM is demonstrated on three different manifolds, arising from three different applications: The sphere, Kendall's shape space (Kendall, 1984), and the space of symmetric, positive definite (SPD) matrices. Furthermore, the WGPLVM is compared with the Euclidean GPLVM, whose predictive distribution is expected not to lie on the manifold. This effect is clearly visible in Fig. 6. A third model, also shown in Fig. 6, is a modification of the Euclidean GPLVM, where the GP predictions are projected onto the manifold in order to make them satisfy the desired constraints.

We first introduce the datasets and their associated tasks, along with dataset-specific details related to training the models. In each case, we train the model assuming independent coordinates, applying the same kernel to each coordinate.

Femur dataset on S^2. A set of directions $P = \{p_i\}_{i=1}^N \in S^2$ of the left *femur* bone of a person walking in a circular pattern (CMU Graphics Lab, 2003; Hauberg, 2016) is measured at $N = 338$ time points. The movement is expected to be one dimensional and periodic, and thus we learn a 1-dimensional submanifold homeomorphic to a circle to approximate the data manifold. The latent variable optimization is initialized using principal curves (Hauberg, 2016), and the prior WGP and GP had kernel

$$k(t, t') = \sigma^2 \exp\left(-\frac{2\sin^2(|t - t'|)/2}{l^2}\right), \qquad (14)$$

and mean $m(t) = \mu_{S^2}$ and $m(t) = 0$, respectively, where μ_{S^2} is the *Fréchet mean* of the training set and σ^2, l^2 are hyperparameters optimized to maximize the likelihood of the dataset P with the latent variables X. The trained models are visualized in Fig. 6.

Diatom shapes in Kendall's shape space. Diatoms are unicellular algae, whose species are related to their shapes. In Kendall's shape space M_K we analyze a set of outline shapes of 780 *diatoms* (du Buf and Bayer, 2002; Jalba et al., 2006) from 37 different species. For visualization, a two dimensional latent space is learned, using the prior $f \sim \mathcal{GP}_{M_K}(m, k)$, with constant basepoint function $m(t) = \mu_{M_K}$ set to be the Fréchet mean of the population and k given by the *radial basis function* (RBF) kernel

$$k(x, x') = \sigma^2 \exp\left(-\frac{\|x - x'\|_2^2}{2l^2}\right). \qquad (15)$$

We initialize the GPLVM and WGPLVM models with PGA and PCA, respectively.

Diffusion tensors in $SPD(3)$. In the space of 3×3 SPD matrices with the Log-Euclidean metric (Arsigny et al., 2006), we collect a set of 750 diffusion tensors from a diffusion MRI dataset, sampled with approximately uniform fractional anisotropy (FA) values. The FA is a well-known tensor shape descriptor; see the supplementary material for the definition. As SPD matrices form an open subset of the Euclidean space of symmetric matrices, *we do not get a "for free" dimensionality reduction* by restricting to SPD matrices. Instead, the data is transformed nonlinearly according to the Log-Euclidean metric, which is commonly used for diffusion tensors (Arsigny et al., 2006). The diffusion MRI image was a single subject from the Human Connectome Project (Glasser et al., 2013; Sotiropoulos et al., 2013; Van Essen et al., 2013). In diffusion MRI, low-dimensional encoding with uncertainty estimates may speed up image acquisition and processing.

Crypto-tensors in $SPD(10)$. On $SPD(10)$ we collect the price of 10 popular crypto-currencies[1] in the time 2.12.2014-15.5.2018. The crypto-currency intra-relationship at a given time is encoded in the covariance matrix between the prices in the past 20 days; we include every 7th day in the period, resulting in 126 10×10 covariance matrices. Wilson and Ghahramani (2011) provide a discussion of covariance descriptors in economy. As the acquired covariance matrices in $SPD(10)$ have eigenvalues in different orders of magnitude, we use the Log-Euclidean metric (Arsigny et al., 2006), capturing this trend better.

For both $SPD(n)$ datasets, the basepoint function, the kernel and the latent variable initialization are chosen as for Kendall's shape space. The latent spaces are chosen to be 2-dimensional for visualization purposes.

Application 1: Encoding. The datasets are divided into training and test sets (consisting of $^8/_{10}$ and $^2/_{10}$ of the data, respectively), and we learn the models φ_{pred} on the training set. Each test element p is "encoded" by the projection $\pi\colon p \mapsto \text{argmax}_{x \in L} \mathbb{P}\{\varphi_{\text{pred}}(x) = p\}$. We assess the quality of this encoding by measuring the root-mean-square error (RMSE) of the reconstruction,

[1] Bitcoin, Dash, Digibyte, Dogecoin, Litecoin, Vertcoin, Stellar, Monero, Ripple, and Verge.

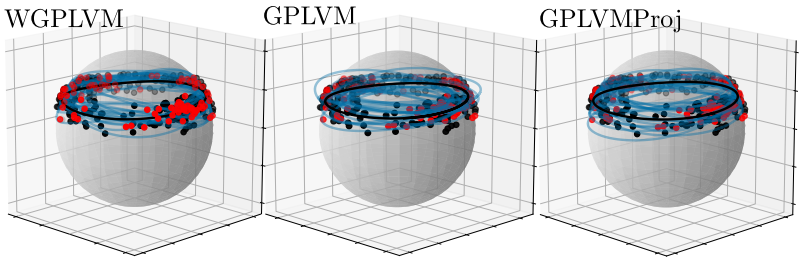

Figure 6: WGPLVM, GPLVM and GPLVMProj submanifold predictions for the *femur* data set. Mean predictions are in black, with 20 samples from the noise models (in blue). Training data in black, with test points in red.

Riemannian	Femur	Diatoms	Diffusion tensors	Crypto-tensors
GPLVMProj	$(9.22 \pm 0.55) \times 10^{-2}$	$(2.48 \pm 0.25) \times 10^{-2}$	0.582 ± 0.025	21.91 ± 2.26
WGPLVM	$\mathbf{(9.20 \pm 0.53) \times 10^{-2}}$	$\mathbf{(2.39 \pm 0.15) \times 10^{-2}}$	$\mathbf{0.391 \pm 0.035}$	$\mathbf{3.04 \pm 0.26}$

Euclidean	Femur	Diatoms	Diffusion tensors	Crypto-tensors
GPLVM	$(9.21 \pm 0.55) \times 10^{-2}$	$(2.48 \pm 0.25) \times 10^{-2}$	$\mathbf{(6.03 \pm 0.34) \times 10^{-2}}$	$(7.36 \pm 5.27) \times 10^{5}$
GPLVMProj	$(9.21 \pm 0.55) \times 10^{-2}$	$(2.48 \pm 0.25 \times 10^{-2}$	$\mathbf{(6.03 \pm 0.34) \times 10^{-2}}$	$(5.49 \pm 3.17) \times 10^{5}$
WGPLVM	$\mathbf{(9.19 \pm 0.53) \times 10^{-2}}$	$\mathbf{(2.39 \pm 0.15) \times 10^{-2}}$	$(7.54 \pm 0.36) \times 10^{-2}$	$(8.69 \pm 7.12) \times 10^{5}$

Table 1: Mean $\pm$ standard error of mean reconstruction errors, measured in RMSE, over 10 repetitions of the experiment. **Top table:** Deviations measured in the intrinsic distance on the manifold. **Bottom table:** Deviations measured in the Euclidean distance.

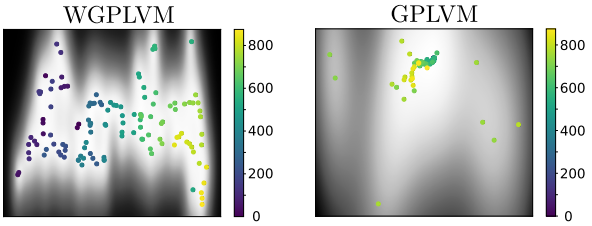

Figure 7: The latent space for the crypto-tensor dataset, with days visualized by color. Note that for GPLVM, the dark blue points corresponding to early times are hidden underneath the green points.

where the error is measured both by the Euclidean metric and the intrinsic metric. Each experiment was repeated 10 times with different training and test sets; the results are reported in Table 1.

Under the intrinsic metric, the WGPLVM performs significantly better on the tensor datasets, and marginally better in the two other cases. Under the Euclidean metric the WGPLVM encoding is better in two cases, worse in one, and inconclusive for the crypto-tensors where no model is significantly better than the others.

Application 2: Uncertainty quantification. Importantly, GPLVM learns a probabilistic model, producing an estimate of uncertainty. We evaluate these uncertainty estimates on all four datasets. Since the predictive distributions live in different spaces, the likelihoods of observed data under the different models are not directly comparable. However, all three models yield confidence intervals, which we compare using 10 resampled training and test sets ($^8/_{10}$ and $^8/_{10}$ of the data). The test set is projected onto the predicted submanifold via π. Then, we sample the respective predictive distributions 50 times, computing the fraction of samples closer to the mean prediction than the test point. The results are visualized in Fig. 8, where the densities of these fractions are shown with corresponding cumulative distributions. For a perfect model fit, we would observe the $x = y$ curve (dashed line) as the cumulative distribution. The experiment shows that all models estimate uncertainty incorrectly, but that WGPLVM obtains the best estimate.

Application 3: Visualization. In Fig. 7, we illustrate the latent spaces of WGPLVM versus GPLVM on the crypto-tensor dataset, which comes with an associated time variable, shown in color. The WGPLVM provides a smoother and more consistent transition in color, while the GPLVM plots all the earlier (dark blue) tensors on top of each other. Similar visualizations for the other datasets can be found in the supplementary material; in these examples, the two visualizations are not significantly different in quality.

In the supplementary material, we provide a discussion on why our model might perform better in the $SPD(n)$ experiments, including a comparison between the Euclidean and Riemannian geometries.

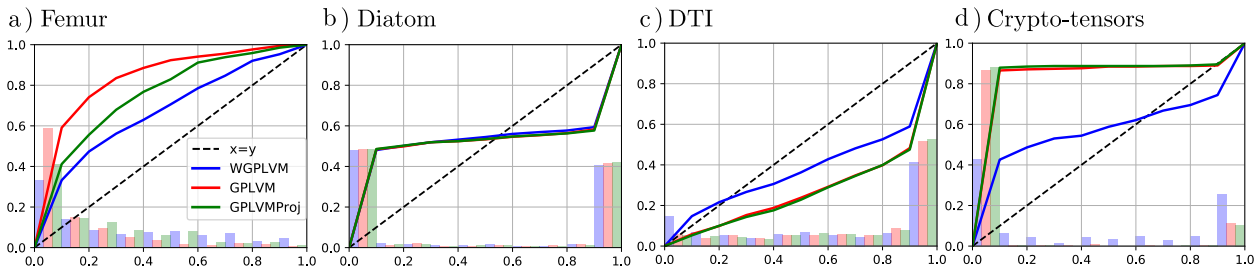

Figure 8: Uncertainty estimates given by the WGPLVM, GPLVM and projected GPLVM models for the four datasets. The bars represent the frequency of occurances, where the fraction of samples, given by the x-value, lie closer to the mean prediction than a test point. The continuous curves represent the cumulative distributions. Whenever the cumulative distribution lies above $x = y$, we are overestimating the corresponding quantile.

5 DISCUSSION AND CONCLUSION

We introduced the WGPLVM for non-parametric and probabilistic submanifold learning on Riemannian manifolds. The model encodes known constraints or invariances, and provides model flexibility, as metrics other than the Euclidean one can be incorporated. This is useful if a different metric captures trends in the data better. The model was evaluated on several manifolds and tasks against the GPLVM and a modified GPLVM, which projects predictions onto the manifold.

The experimental results show that the WGPLVM provides a better probabilistic model to fit the data; in particular the uncertainty estimates are superior to the Euclidean models on three out of four datasets, and virtually identical on the fourth. We note that for Euclidean models, the uncertainty is visibly higher. These are strong indications that our model carries out modelling the data distribution better. The mean predictions of the WGPLVM encode the data space significantly better than the GPLVM and projected GPLVM models on two of the datasets, and marginally better on the other two, when measured in the Riemannian metric. Under the Euclidean metricr, the GPVLM performs notably better in one experiment, and WGPLVM marginally better in two. On cryptotensors, we deem the results inconclusive due to high variance. The aforementioned effects are also seen in the latent space visualizations, e.g. on the cryptotensors the WGPLVM better detects small-scale differences in the early time steps.

One might suspect that the improved performance stems from a "for free" dimensionality reduction through constraints. However, we note that the most significant improvement in both reconstruction error and visualization was obtained on $SPD(n)$, where the Riemannian manifold is a full-dimensional, convex subset of the Euclidean ambient space. This might still be due to the constraints, which forces the distributions to lie in the manifold. The difference could also be caused by the choice of metric. For the crypto-tensors in particular, we observe that some of the eigenvalues are very small; the Log-Euclidean metric essentially acts as a log-transform and therefore converts the data to a scale on which changes in the smaller eigenvalues can be detected.

In three of the experiments, the mean predictions of GPLVM lie essentially on the manifold, thus the projected version does not improve the mean reconstruction error. However, in the femur experiment, the uncertainty estimates are clearly improved, but also notably outperformed by WGPLVM. Due to the metric and curvature of the manifold, interpolation between two points in the ambient space $\mathbb{R}^n$ does not necessarily project even closely onto the manifold interpolation between the projected points. This distortion affects the statistics relying on interpolation, and explains both the reduced reconstruction capability and the increased variance. Furthermore, the projected model ignores any metric choices imposed on the manifold.

Although the WGPLVM provides flexibility through the prior basepoint function, we fixed this to be the Fréchet mean of the training set in our experiments. The choice is well justified if the data is local enough, and makes the comparison to GPLVM fair. The flexibility to delocalize the learning process through the basepoint function is, however, important for inference on manifolds when the locality assumption fails. The non-trivial optimization of the basepoint function thus provides a venue for future research.

In summary, the WGPLVM is a probabilistic submanifold learning algorithm that respects known Riemannian manifold structure in the data by taking values in the associated Riemannian manifold. We compare the model to its Euclidean counterparts on a number of manifolds, datasets and tasks, and show that it has superior representation capabilities more faithful visualizations and improved uncertainty estimates.

Acknowledgements

AM and AF were supported by Centre for Stochastic Geometry and Advanced Bioimaging, funded by a grant from the Villum Foundation, and SH was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 757360), as well as a research grant (15334) from VILLUM FONDEN. Data were provided [in part] by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. The data [in part] used in this project was obtained from `mocap.cs.cmu.edu`. The database was created with funding from NSF EIA-0196217. The authors wish to thank Thomas Hamelryck for helpful comments.

References

Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Log-euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 56(2):411–421, 2006.

Phillipp G. Batchelor, Maher Moakher, David. Atkinson, Fernando. Calamante, and Alan Connelly. A rigorous framework for diffusion tensor calculus. *Magnetic Resonance in Medicine*, 53(1):221–225, 2005. ISSN 1522-2594.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

CMU Graphics Lab. CMU Graphics Lab Motion Capture Database . `http://mocap.cs.cmu.edu/`, 2003. The data used in this project was obtained from mocap.cs.cmu.edu. The database was created with funding from NSF EIA-0196217.

Manfredo Perdigao Do Carmo. *Riemannian geometry*. Birkhauser, 1992.

Hans du Buf and Micha Bayer. *Automatic Diatom Identification*. 2002.

P. Thomas Fletcher and Sarang Joshi. *Principal Geodesic Analysis on Symmetric Spaces: Statistics of Diffusion Tensors*, pages 87–98. 2004.

P. Thomas Fletcher, Conglin Lu, Stephen M. Pizer, and Sarang Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE transactions on medical imaging*, 23(8):995–1005, 2004.

Oren Freifeld and Michael J Black. Lie bodies: A manifold representation of 3D human shape. In A. Fitzgibbon et al. (Eds.), editor, *European Conference on Computer Vision (ECCV)*, Part I, LNCS 7572, pages 1–14. Springer-Verlag, 2012.

Matthew F. Glasser, Stamatios N. Sotiropoulos, J. Anthony Wilson, Timothy S. Coalson, Bruce Fischl, Jesper L. Andersson, Junqian Xu, Saad Jbabdi, Matthew Webster, Jonathan R. Polimeni, et al. The minimal preprocessing pipelines for the Human Connectome project. *Neuroimage*, 80:105–124, 2013.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

Søren Hauberg. Principal curves on Riemannian manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2016.

Søren Hauberg, Oren Freifeld, and Michael J. Black. A geometric take on metric learning. In P. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems (NIPS) 25*, pages 2033–2041. MIT Press, 2012.

Stephan Huckemann, Thomas Hotz, and Axel Munk. Intrinsic shape analysis: Geodesic PCA for Riemannian manifolds modulo isometric Lie group actions. *Statistica Sinica*, pages 1–58, 2010.

Andrei C. Jalba, Michael HF Wilkinson, and Jos BTM Roerdink. Shape representation and recognition through morphological curvature scale spaces. *IEEE Transactions on Image Processing*, 15(2):331–341, feb 2006.

David G. Kendall. Shape manifolds, Procrustean metrics, and complex projective spaces. *Bulletin of the London Mathematical Society*, 16(2):81–121, 1984.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, 2014.

Neil D Lawrence. Gaussian process latent variable models for visualisation of high dimensional data. In *Advances in neural information processing systems*, pages 329–336, 2004.

Neil D. Lawrence. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. *J. Mach. Learn. Res.*, 6:1783–1816, 2005. ISSN 1532-4435.

Neil D. Lawrence and Andrew J. Moore. Hierarchical Gaussian process latent variable models. In *Proceedings of the 24th international conference on Machine learning*, pages 481–488. ACM, 2007.

Lizhen Lin, Mu Niu, Pokman Cheung, and David Dunson. Extrinsic gaussian processes for regression and classification on manifolds. *arXiv preprint arXiv:1706.08757*, 2017.

Anton Mallasto and Aasa Feragen. Wrapped Gaussian process regression on Riemannian manifolds. In *CVPR - IEEE Conference on Computer Vision and Pattern Recognition, to appear*, 2018.

Kanti V. Mardia and Peter E. Jupp. *Directional statistics*, volume 494. John Wiley & Sons, 2009.

Xavier Pennec. Barycentric subspaces and affine spans in manifolds. In *Geometric Science of Information - Second International Conference, GSI 2015, Palaiseau, France, October 28-30, 2015, Proceedings*, pages 12–21, 2015.

Xavier Pennec. Barycentric subspace analysis on manifolds. *arXiv preprint arXiv:1607.02833*, 2016.

Carl Edward Rasmussen. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pages 63–71. Springer, 2004.

Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *SCIENCE*, 290:2323–2326, 2000.

Stamatios N. Sotiropoulos, Steen Moeller, Saad Jbabdi, Jungqian Xu, Jesper Andersson, Edward John Auerbach, Essa Yacoub, David A. Feinberg, Kawin Setsompop, Lawrence L. Wald, et al. Effects of image reconstruction on fiber orientation mapping from multichannel diffusion MRI: reducing the noise floor using SENSE. *Magnetic resonance in medicine*, 70 (6):1682–1689, 2013.

Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

Michalis Titsias and Neil D. Lawrence. Bayesian Gaussian process latent variable model. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 844–851, 2010.

Oncel. Tuzel, Fatih Porikli, and Peter Meer. Region covariance: A fast descriptor for detection and classification. *European Conference on Computer Vision (ECCV)*, pages 589–600, 2006.

Raquel Urtasun and Trevor Darrell. Discriminative Gaussian process latent variable model for classification. In *Proceedings of the 24th international conference on Machine learning*, pages 927–934. ACM, 2007.

Raquel Urtasun, David J. Fleet, and Pascal Fua. 3d people tracking with Gaussian process dynamical models. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 238–245. IEEE, 2006.

Raquel Urtasun, David J. Fleet, Andreas Geiger, Jovan Popović, Trevor J. Darrell, and Neil D Lawrence. Topologically-constrained latent variable models. In *Proceedings of the 25th international conference on Machine learning*, pages 1080–1087. ACM, 2008.

David C. Van Essen, Stephen M. Smith, Deanna M. Barch, Timothy EJ Behrens, Essa Yacoub, Kamil Ugurbil, Wu-Minn HCP Consortium, et al. The wu-minn Human Connectome project: an overview. *Neuroimage*, 80:62–79, 2013.

Aydin Varol, Mathieu Salzmann, Pascal Fua, and Raquel Urtasun. A constrained latent variable model. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2248–2255. Ieee, 2012.

Andrew Gordon Wilson and Zoubin Ghahramani. Generalised Wishart processes. In *UAI 2011, Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, Barcelona, Spain, July 14-17, 2011*, pages 736–744, 2011.

Miaomiao Zhang and P. Thomas Fletcher. Probabilistic principal geodesic analysis. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 1178–1186, 2013.